Prethodno priopćenje UDK [130.1:159.9.016.1]:004.8
Primljeno 31. 3. 2015.

Sandro Skansi
Bolnička 34f, HR–10090 Zagreb
sandro.skansi@infigo.hr

Umjetna inteligencija i kompatibilizam: mogućnost postanka slobodnog uma u determiniranom tijelu

Sažetak

Ovaj rad istražuje mogućnost davanja kompatibilističkog argumenta iz aspekta umjetne inteligencije. Ključna pretpostavka našeg rada jest da je umjetna inteligencija načelno moguća i da se realizira na računalnim arhitekturama u bitnome nalik današnjim. Uz taj je uvjet moguće dati definiciju slobode koja je pomirljiva s determiniranim izračunom, uz pomoć načelne nedokučivosti inteligentnog procesa. Ovo se temeljem funkcionalizma može translatirati u filozofiju (ljudskog) uma. Pitanje je li moguće naš argument adaptirati za drugačije teorije filozofije uma ostavljamo otvorenim.

Ključne riječi

umjetna inteligencija, nedeterminizam, funkcionalizam, kompatibilizam, razlikovanje um–tijelo, slobodan izbor

Uvod

Pitanje odnosa uma i tijela jedno je od temeljnih filozofijskih pitanja. Ovo je pitanje koje poprima razne oblike i dobiva raznovrsne argumente kao odgovore. Mnogi od ovih argumenata mehanicističke su prirode, tako da pokušavaju objasniti »zašto« objašnjavajući »kako«. Ova je kategorijalna pogreška[1] možda jedna od najsretnijih instanci kategorijalne pogreške jer zahvaljujući njoj danas imamo mnoge modele ove veze i jasno argumentirane i raznolike načine povezivanja uma i tijela.

McCarthy i Hayes su još kasnih 1960-ih (McCarthy, Hayes 1969) u svom radu argumentirali da je od nekoliko reprezentacija prirodnog svijeta za potrebe umjetne inteligencije samo nekoliko stvarno primjenjivo u kontekstu računala. Mi se u ovom radu želimo nadovezati na tu tradiciju, ali želimo nadograditi ovu teoriju i zanima nas, s jedne strane, kakvu filozofiju »umjetnog« uma možemo izgraditi, kao i kakvu teoriju u filozofiji (ljudskog) uma povlači pretpostavka da imamo gotovu umjetnu inteligenciju. Prema pretpostavci za smjer »s desna na lijevo«, prava umjetna inteligencija je prihvaćeno inteligentna i bazirana je na računalnom sklopovlju.

Iz ovog razloga, za svaku teoriju koja nije utemeljena na ideji tijela (mozga) kao supstrata neke vrste nad kojim se realizira um nije adekvatna već na sa-

[1] Pojam kategorijalne pogreške je standardan, premda može biti višeznačan, ukoliko se ne uzima u strogom smislu filozofije uma. Mi ga uzimamo u izvornom značenju (Ryle 1949).

mom početku jer se umjetna inteligencija realizira nad računalnim sklopovljem.

Jedna od najintrigantnijih teorija, koju ćemo u ovom radu zastupati, jest kompatibilizam koji pokušava pomiriti determinizam prirode i slobodu uma. Pokazat ćemo da se upravo kroz umjetnu inteligenciju ovaj stav može opravdati te da bi bilo prikladno simulirati slobodnog agenta koji koristi deterministično sklopovlje upravo realizacijom kompatibilističke teorije.

Kompatibilizam se tradicionalno zastupao na drugačiji način. Primjerice, Goodwin prenosi da je Carnap tvrdio (Goodwin 2009: 11–12) da je sloboda izbora predicirana nad determinizmom jer, da bismo imali izbor koji je slobodan, a ne nasumičan, moramo imati poimanje o tome da će se stvari razvijati barem u doglednoj budućnosti s obzirom na odabir, što pretpostavlja ne samo determinirani lanac događaja nego i činjenicu da je spoznatljivo determiniran. Ovdje je važno istaknuti da za kompatibiliste slobodna volja ne implicira slobodu djelovanja (usp. Dennet 1984; Frankfurt 1971) – premda možemo odlučiti nešto napraviti, npr. poletjeti, to ne znači da možemo to stvarno i napraviti ako se suprotstavlja zakonima fizike, ili u nekim slučajevima zakonima društva koji su često u krajnoj instanci poduprti zakonima prirode (npr. oružjem, zatvorom). Ovo će se nadovezati na našu diskusiju oko verzijskog prostora pa nam je zato posebno relevantno.

Neki pristupi u umjetnoj inteligenciji, poput neuralnih mreža,[2] prvotno su bili utemeljeni na teoriji identiteta[3] (Smart 1956; Place 1956) koja tvrdi da je mentalno stanje funkcijski korelirano sa stanjem mozga. Preciznije rečeno, za svako je stanje mozga moguće naći mentalno stanje koje mu odgovara i obrnuto. Glavni prigovor ovom pristupu je takozvana višestruka ostvarivost, odnosno činjenica da se isto (ili neraspoznatljivo slično) mentalno stanje može pobuditi vrlo različitim stanjem mozga. Jednostavan primjer toga su npr. vizualne i olfaktorne asocijacije, koje daju jednak učinak u smislu mentalnog stanja koje se pobuđuje, ali (neurološki gledano) radikalno drugačijim putem. Neurološki detalji su zanimljivi, ali – što zbog naše nedovoljne stručnosti, što zbog toga što to izlazi iz okvira teme – ne ulazimo dublje u ovu raspravu.

Bez obzira na to što su neuralne mreže potekle iz teorije identiteta (koja u svakom slučaju ima ozbiljnih poteškoća), njihova je tehnološka primjenivost iznimna. Dio razloga zašto je tomu tako jest upravo višestruka ostvarivost jer je moguće argumentirati da je dvjema realizacijama na različitim procesima mozga (u ovom slučaju na prirodnom i umjetnom mozgu – jedan neuralni, a drugi računalni) moguće dobiti isto mentalno stanje. Naglasak nije na tome da je to činjenično stanje stvari nego da isti argument, koji poništava jednostavnu teoriju identiteta, široko otvara vrata umjetnoj inteligenciji.

Veza slobode i (umjetne) inteligencije

Jedna provizorna definicija slobode, odnosno slobodne volje, koju želimo usvojiti sadržava sljedeće značajke:

1. Slobodan čin je ili *prima facie* racionalan ili *post facto* racionalan, no u određenom je broju slučajeva nasumičan u nekom verzijskom prostoru, a u određenom je broju slučajeva nasumičan u najopćenitijem verzijskom prostoru.
2. Racionalan čin je čin utemeljen na znanju zatvorenom za njegove logičke posljedice koji je zatvoren pod nekim (socijalno) prihvatljivim sustavom zaključivanja.

Ovdje trebamo napomenuti da smo pojam 'racionalan' uzeli u značenju koje isključuje mogućnost neobrazloženog eksperimenta kao racionalnog. Neobrazložen eksperiment u nedostatku rješenja jest racionalan odabir, no socijalne norme ovo ne uključuju odmah kao racionalno, nego svoje priznanje dobiva tek *post facto*.

Naš razlog relativno je rudimentarniji, a to je da nam treba pojam za »epistemički utemeljen i zatvoren na posljedice socijalno prihvatljivog sustava zaključivanja«, pa za to koristimo pojam 'racionalan'. Još nam je potrebna definicija verzijskog prostora. Verzijski prostor je prostor odabira načina djelovanja i može biti ili fizikalno uvjetovan ili socijalno uvjetovan. Kada se kaže »socijalno uvjetovan«, to nije u smislu »mekih« socijalnih normi poput pristojnosti, nego »tvrdih« normi poput prava prednosti u prometu ili zatvora, što dovodi direktno do fizikalno uvjetovanog (ako je agent zatvoren, ne može izaći iz zatvora zbog fizikalnih zakona).

Primjer odluke unutar i izvan verzijskog prostora najbolje je ilustrirati time da bilo koji agent može odlučiti pasti s nekog povišenog mjesta, ali samo neki agenti mogu odlučiti poletjeti. Radi jednostavnosti, zamislimo da su jedini mogući agenti ptice i miševi te da su oni u dovoljnoj mjeri svjesni da su u stanju odlučiti između više načina djelovanja da dođu do svog cilja. Ptica može odlučiti pasti (što temeljem gravitacije mogu svi agenti i taj način djelovanja zato pripada najopćenitijem verzijskom prostoru) ili poletjeti (što je specifičan način djelovanja za ptice). Miš, s druge strane, može samo odlučiti pasti jer on u svom verzijskom prostoru nema mogućnost letenja.

Iz ovoga bi se dalo zaključiti da je verzijski prostor ustvari prostor djelovanja, ali on je općenitiji od toga jer se tiče bilo koje odluke, uključujući načine razmišljanja ili zaključivanja. Ako, primjerice, imamo problem koji je zadan s A v B v C, mi ga možemo pokušati riješiti s A, B ili C. Ovdje nam »nađi rješenje problema zadanog s X« znači »X je istinit i nađi takav Y da Y→X je istinito«. Ako nemamo dodatnih podataka, rješenje za A v B v C može biti i A i B i C (koji tvore verzijski prostor), no problem je u tome što nam najčešće ne treba jedno rješenje, nego kada odaberemo npr. rješenje A, onda idemo dalje tim smjerom. Ukoliko taj smjer uđe u slijepu ulicu, moramo ići natrag do prethodnog križanja i pokušati drugom ulicom (ovo se naziva *backtracking*). Zbog toga A, B i C nisu nužno jednakovrijedni izbori, nego smislenost odabira ovisi o tome što će slijediti.

Dodatna se poteškoća javlja ako naše stablo izbora ima neke nagrade. Npr. ako odaberemo A dobijemo 5, ako odaberemo B dobijemo 7 i ako odaberemo C dobijemo 10. Ako ne znamo što ide dalje, smisleno bi bilo uzeti C, no ako u sljedećem koraku imamo, ako smo uzeli C, izbor između CA (nagrada 2) i CB (nagrada 1), a iza A imamo AA (nagrada 8) i AB (nagrada 9), bez obzira na to što ćemo (biti prisiljeni) odabrati u sljedećem koraku, više se isplati odabrati A u prvom jer je minimalna kumulativna nagrada za sljedeći korak u smjeru A veća od maksimalne kumulativne nagrade za sljedeći korak u smjeru C. Algoritam koji na svakom koraku uzima samo maksimalno što se nudi odmah zove se »pohlepnim« i on je vrlo jednostavan i zahtijeva jako malo zaključivanja i planiranja.[4]

2
Za detalje upućujemo na povijesno važnu knjigu: Minsky i Papert 1987, a za suvremeni pristup i moderne metode koje se danas koriste upućujemo na: Russell i Norvig 2009.

3
Teoriju identiteta uzimamo kako je opisana u: Place 1956 i Smart 1956.

4
Svakako je zanimljivo ovo povezati s pohlepnošću kod ljudi i sposobnošću pohlepnih da dođu do optimalne dugoročne dobiti, no ta bi nas diskusija odvela predaleko od naše teme.

Ako su ovi koraci postavljeni tako da idu jedan neposredno iza drugog, tada je u kontekstu vrijednosti nagrade 1 + 1 = 2. Ako se pak treba čekati, utilitarna je vrijednost neposredne nagrade veća od dalje nagrade, pa je 1(neposredna) + 2(dalja) < 2(neposredna) + 1(dalja), možda čak i 1(neposredna) + 3(dalja) < 2(neposredna) + 1(dalja) ili 1(neposredna) + 10(dalja) < 2(neposredna) + 1(dalja).

Ako dodamo činjenicu da je u smjeru A sljedeća prilika za odluku udaljena 10 dana, a u smjeru C udaljena 5 godina, izbor postaje još složeniji. Dodatna je komplikacija ako ovi intervali nisu unaprijed definirani. Posebno ako se iz smjera A izbori iz C više ne vide direktno nego izobličeno (čak i nasumično izobličeno). Očito je da je ovo analiza procesa planiranja kako se on manifestira u svijetu i služit će nam kao mjera slobode s kojom ćemo uspoređivati naš model. Detalji nisu bitni, ali ono što treba uočiti jest da se mnogi procesi mogu prikazati numerički (dodavanjem nekog faktora, pondera), ali se *backtracking* ne može provesti, kao što se »magla« koja prekriva buduće izbore ne može anulirati doradom od strane agenta.

U (1.) tvrdimo da je slobodan čin ili (1.a) *prima facie* racionalan ili (1.b) *post facto* racionalan, ili (1.c) nasumičan u nekom određenom verzijskom prostoru ili (1.d) nasumičan u najopćenitijem verzijskom prostoru.

Ako je (1.a), onda je racionalan prema (2.), što znači da je baziran na nekom epistemički utemeljenom stavu (istinitom, opravdanom vjerovanju) koji je zatvoren pod nekim socijalno prihvatljivim sustavom zaključivanja. Pod »socijalno prihvatljiv sustav zaključivanja« podrazumijevamo mogući logički sustav koji je društvo sankcioniralo kao ispravan. Iz te perspektive sustav propozicionalne logike ne bi bio (u potpunosti) ispravan zbog implikacije u slučaju kada je antecedent neistinit, ali bi sustav silogistike bio ispravan, kao i neki sustavi neklasičnih logika.

Kao što vidimo, veliki je naglasak stavljen na socijalni aspekt sustava zaključivanja, što bi impliciralo da je slobodan izbor barem djelomično socijalni konstrukt. Ako gledamo eksternalistički (kada ćemo neku odluku nazvati slobodnom), to je sasvim razumljivo, ali mi zauzimamo jači stav, slijedeći Wittgensteinove argumente (Wittgenstein 1953): nije moguće dati bilo kakav spoznatljiv sadržaj pojmu ‘slobodan izbor’ osim temeljem jezičnog okvira koji se internalizira, pa je stoga kognitivni sadržaj osobnog shvaćanja vlastitog slobodnog izbora isto što i shvaćanje tuđeg slobodnog izbora i svodi se na to kojim drugim jezičnim faktorima zajednica objašnjava pojam slobode.

Ako je (1.b), onda vrijede iste napomene, ali je situacija malo kompliciranija. Ovdje slobodni agent djeluje prema svojoj prosudbi, ali je smjer djelovanja tek temeljem rezultata socijalno opravdan. Razlog je to što se već internalizirana norma socijalno ispravnog zaključivanja instancira u agentovom umu na način koji nije brzo dokučiv drugim agentima. Ovaj slučaj obuhvaća naše temeljno shvaćanje inteligencije kao racionalne vrline, odnosno sposobnosti da se nađe rješenje koje drugima nije palo na pamet. Ovdje je vidljiva ključna veza između slobodnog izbora i inteligencije, a vidljivo je i da je potrebno jedno razmatrati kroz aspekt drugoga.

Aspekti (1.c) i (1.d) tiču se nasumičnih izbora unutar specifičnog verzijskog prostora (1.c) i najopćenitijeg verzijskog prostora (1.d). Kako je najopćenitiji verzijski prostor prostor svih mogućnosti, ovo je u potpunosti nasumičan odabir odluke. (1.c) je specijalniji, ali i dalje nasumičan odabir u verzijskom prostoru odabira koji su smisleni u nekom kontekstu.

Zanimljivo je vidjeti da se, s jedne strane, slobodni izbor sastoji od (1.a) ili (1.b) ili (1.c) ili (1.d) (zamislimo da će onda vjerojatnost da slobodni agent

koristi pojedini slučaj biti 25 %), dok bi se inteligentni izbor mogao definirati, primjerice, kao: u 50 % slučajeva koristi (1.a), u 30 % slučajeva (1.b), u 15 % slučajeva (1.c) i u 5 % slučajeva (1.d). Ovo govori u prilog tome da je glavna razlika između slobodnog odabira i inteligentnog odabira distribucija postotaka, odnosno da je prijelaz od imanja slobode za izbor do inteligetnog izbora samo pitanje namještanja postotaka ili, drugim riječima, da je čovjek rođen s mogućnošću slobodnog izbora unaprijed postojećeg u umu, a učenje namješta ove postotke. Time se također može obrazložiti zašto su djeca i mladi skloniji eksperimentalnom ponašanju, posebno unutar najopćenitijeg verzijskog prostora.

U izradi umjetne inteligencije postoje nekompetitivni (npr. slaganje slagalice), kompetitivni (npr. igranje šaha) i semikompetitivni (npr. vožnja auta) konteksti. Korištenje nasumičnosti (posebno u najopćenitijem verzijskom prostoru), osim ekperimentiranja, korisna je tehnika zbunjivanja protivnika u kompetitivnim kontekstima, pa će tamo biti češće upotrebljavana.

Kompatibilizam i slobodna funkcija

Da bismo opisali slobodu u kontekstu računalnog programa, želimo definirati ono što ćemo nazvati »slobodnom funkcijom«. Da bismo izgradili kompatibilistički argument, našu teoriju uma koncipiramo kao nadogradnju nad funkcionalizmom (Putnam 1960),[5] što je prirodan izbor s obzirom na to da govorimo o umjetnoj inteligenciji. Ovdje će nam biti presudno razlikovati dva tipa determinizma jer se nezgodno jednako zovu, ali se referiraju na vrlo različite fenomene. Prvi je fizikalni determinizam, odnosno determinizam prirode (često shvaćen kao kontraran slobodnoj volji), a drugi je računalni determinizam.

Računalni determinizam (i nedeterminizam) opisuje postupak izračuna (usp. Sipser 2012). Deterministički je izračun onaj koji ima funkcionalni prijelaz iz stanja u stanje, što znači da se točno zna iz kojeg će stanja prijeći u koje, dok se u slučaju nedeterminizma prijelaz iz stanja u stanje definira relacijski. Ovo je smisleno, no u praktičnom se smislu može postići jedino tako da se izračun masivno paralelizira i da se svaki od tih procesa paralelno ostvaruje u memoriji. Problemi povezani s time su: (a) možda ne znamo točno koliko nam se proces grana (ovo djeluje kao tehnikalija, no u stvari je od ključne važnosti), a memoriju moramo dodati na početku, i posebno (b) broj paralelnih izračuna može biti iznimno velik.

Za problem, kada ga prikažemo kao funkciju odlučivanja (koja daje 0 ili 1, s jasnim tumačenjem), koji deterministički Turingov stroj (DTS) rješava u polinomijalnom vremenu, kažemo da spada u klasu P, a ako ga u polinomijalnom vremenu može riješiti tek nedeterministički Turingov stroj u polinomijalnom vremenu (NTS), kažemo da spada u klasu NP. Vjeruje se, ali je nedokazano, da ove klase nisu iste.

Rješenje problema koji spada u klasu NP, *deterministički* Turingov stroj može *provjeriti* u polinomijalnom vremenu. Temeljem ovoga želimo »*recastati*« definiciju slobode u kontekstu potencijalne kompatibilističke teorije. Očito je da želimo igrati u smjeru (1.b), jer time imamo mogućnost koja je *post facto* opravdana (deterministički je stroj provjerava u doglednom vremenu), ali je ne može izračunati u doglednom vremenu. (1.a) možemo predstaviti determi-

5
Klasični početak funkcionalizma opisan je u: Putnam 1960. Opširniji opis koji uključuje i razvoj funkcionalizma može se naći u: Shagrir 2005.

nističkim izračunom (nedeterministički Turingov stroj može deterministički računati), (1.c) i (1.d) su *de facto* nasumični odabiri (razlika između ova dva odabira je važna, ali ne za naš sadašnji cilj).

Odabir (1.a) deterministički je izračun koji drugi deterministički stroj gleda i provjerava. (1.b) bi predstavljao nedeterministički izračun koji neki drugi (deterministički) stroj gleda i provjerava. Ovo u općem slučaju može napraviti samo nedeterministički Turingov stroj, no u posebnim slučajevima to se može dogoditi i s determinističkim strojem.

Zamislimo da dva DTS-a računaju neki problem, primjerice, traže lozinku do duljine 10, i jedan počinje s 0000000000 i povećava, a drugi s 9999999999 i smanjuje. Ako je lozinka 9999954321, onda će je drugi DTS mnogo ranije pronaći, i zamislimo da, kada je pronađe, javi drugom (koji još računa) »gotov sam« i onda je drugi provjeri. Ovo je pojednostavljenje, ali pokazuje da temeljem početne konfiguracije (slučajnošću) jedan agent može biti bitno brži od drugog.

No moguće je i bolje s DTS-om, da ne ovisi o slučajnosti. Ideja se naziva heuristikom: možda je moguće pronaći dobre aproksimacije za NP-teške probleme[6] koji su polinomijalni. Ovo je često moguće napraviti strojnim učenjem, koje smanjuje verzijski prostor nemogućeg i testiraju rezultat. No to izlazi izvan okvira ovog rada.

Zamislimo dva DTS-a (DTS1 i DTS2) koji računaju neki NP-problem. DTS1 ima prikladnu heuristiku koja je strojnim učenjem dobro naučena za određeni problem i tada ga DTS1 brže rješava. DTS2 ga provjerava i verificira da je rješenje dobro. Iz perspektive DTS2, DTS1 je dobro riješio, samostalno ga je riješio i ovo dobro rješenje ovisi o uvježbanosti. Po svakom kriteriju, ovo je racionalan postupak. Ali važnije od toga je da je postupak slobodan u smislu da je DTS2-ovo rješenje lakše opisati fazu po fazu nego opisati funkciju koja to rješava (koju DTS1 primjenjuje zato što ju je naučio), što znači da je, usprkos determinističkom izračunu, prikladan opis DTS1 onaj koji ga opisuje kao da slobodno postupa u tom zadatku.

Ovo nije pitanje jezika nasuprot stvarnosti ili jezičnoj jednostavnosti, nego za DTS2 načelno nije pojmljiva deterministička funkcija za koju bi DTS1 računao. Ovo je suptilna razlika jer se radi o tome da DTS2 mora prihvatiti i determiniranost i slobodu. Kada bi DTS2 naučio tu funkciju, ili čak i sve slične funkcije, DTS1 (ili neki treći) na isti bi način to učinio za neke kompleksnije funkcije i tako do limita izračunljivosti.

Ovo nije toliko egzotično koliko zvuči jer se programe iz umjetne inteligencije upravo razvija na taj način i čest je pristup, kada se funkcija jednom objasni, »ma, to nije inteligencija (sloboda), to je samo izračun«. Ovo više govori o našim očekivanjima, jer ako bismo dosljedno proveli taj stav, za svaki bi program umjetne inteligencije to značilo da umjetna inteligencija načelno nije moguća, što je legitiman stav, ali nas taj stav ostavlja na *statusu quo*.

Ovo bi značilo da ne želimo priznati mogućnost inteligencije (pa ni slobode) stroju, što manje govori o realnosti (koja bi se u ovom slučaju iz jasnih razloga sastojala isključivo od emprijske realnosti spoznatljivih svojstava), a više o našem doživljavanju inteligencije kao inherentno ljudske kvalitete. Ako ćemo odbaciti svaki algoritam kao »samo izračun«, bez obzira na to koliko dobro imitira ljude (mnogi prolaze Turingov test, kao jedan od glavnih kriterija), tada mi *ante rerum* zaustavljamo mogućnost da se ikada pojavi izračun (uvijek će na računalu to biti »samo« neki izračun) koji je inteligentan. No ako priznajemo da je umjetna inteligencija načelno ostvariva, slijedi kompatibilistički argument za strojeve, što preko funkcionalizma postaje kompatibilističkim argumentom za ljudski um.

Zaključak

U ovom smo radu istražili mogućnost kompatibilističkog argumenta za filozofiju (ljudskog) uma. Vjerujemo da je moguće argumentirati o umu kao nečemu inheretno ljudskom, no mi nismo htjeli prihvatiti ovaj stav jer je dojam da se odvajujući neki proces ili svojstvo kao ljudsko gubi esencijalno svojstvo koje ga odvaja. Ako je to jedino esencijalno svojstvo razlike (u što ubrajamo, naravno, i njegova derivirana svojstva), tada u svakom spoznatljivom smislu razlika ne postoji. Ovo bismo bili spremni nazvati »inkluzivnim« gledištem. Temeljem inkluzivnog pogleda na umjetnu inteligenciju možemo pretpostaviti da je umjetna inteligencija moguća, u značenju da je moguće stvoriti strojeve koji su u svom djelovanju dosljedni istoj ideji inteligencije kojoj i ljudska bića teže,[7] što znači da je isti argument načelno prenosiv između umjetnih umova i ljudskih umova. Kao jedan od temeljnih argumenata protiv umjetnog shvaćanja slobode navodimo *de facto* subjektivnu percepciju vrijednosti, odnosno: kada bi umjetni um mogao donijeti slobodnu odluku, bio bi umnogome sličniji ljudskom umu.

Literatura

Dennett, D. 1984. *Elbow Room: The Varieties of Free Will Worth Wanting*. London: Bradford Books.

Frankfurt, H. 1971. Freedom of the Will and the Concept of the Person. *Journal of Philosophy*, god. 68, sv. 1, str. 5–20.

Goodwin, C. J. 2009. *Research In Psychology: Methods and Design*. New York: Wiley.

McCarthy, J.; Hayes, P. J. 1969. Some Philosophical Problems from the Standpoint of Artificial Intelligence. U: *Machine Intelligence*. Ur. B. Meltzer, D. Michie. Edinburgh: Edinburgh University Press, str. 463–502.

Minsky, M.; Papert, S. 1987. *Perceptrons: An Introduction to Computational Geometry*. Cambridge: MIT Press.

Place, U. 1956. Is Consciousness a Brain Process? *British Journal of Psychology*, god. 47, sv. 1, str. 44–50.

Putnam, H. 1960. Minds and Machines. U: *Mind, Language, and Reality.* Cambridge: Cambridge University Press, str. 362–385.

Russell, S.; Norvig, P. 2009. *Artificial Intelligence: A Modern Approach*. Harlow: Prentice Hall.

Ryle, G. 1949. *The Concept of Mind*. London: Hutcheson.

Shagrir, O. 2005. The Rise and Fall of Computational Functionalism. U: *Hilary Putnam*. Ur. Y. Ben-Menahem. Cambridge: Cambridge University Press, str. 220–250.

Sipser, M. 2012. *Introduction to the Theory of Computation*. Boston: Cengage Learning.

Smart, J. J. C. 1956. Sensations and Brain Processes. *Philosophical Review*, god. 68, sv. 2, str. 141–156.

Wittgenstein, L. 1953. *Philosophical Investigations*. Oxford: Blackwell.

6
NP-teški problemi su problemi koji su u NP u prosječnom, ne u najgorem slučaju. Za njih se također (još i više) koriste heuristike, ali osnovni je princip isti kao i kod NP-problema.

7
Kako zauzimamo inkluzivnu poziciju, ne smijemo pretpostaviti da stroj slijedi ljudsku inteligenciju, nego da slijedi inteligenciju kao isti ideal koji i ljudi slijede u razvoju svoje inteligencije.

Sandro Skansi

Artificial Intelligence and Compatibilism: Possibility of Emergence of the Free Mind in the Determined Body

Abstract

This paper explores the possibility of a compatibilistic argument from the aspect of artificial intelligence. A key assumption for our argument is that artificial intelligence is in principle possible and that it is realized on computer architectures similar to today's architectures. With these assumptions, it is possible to give a definition of freedom which is compatible with a deterministic calculation, by using unattainableness of intelligent process computation. By using functionalism as a background theory, this can be translated in philosophy of (the human) mind. The question whether our argument is adaptable to different theories in philosophy of mind is left open.

Key words

artificial intelligence, non-determinism, functionalism, compatibilism, mind and brain distinction, free choice